Review

Cell PRESS

# Molecular evolution of the RNA polymerase II CTD

**Rob D. Chapman, Martin Heidemann, Corinna Hintermair and Dirk Eick**

Institute for Clinical Molecular Biology and Tumour Genetics, Helmholtz Center for Environmental Health, Center for Integrated Protein Science (CiPSM), D-81377 Munich, Germany

**In higher eukaryotes, an unusual C-terminal domain (CTD) is crucial to the function of RNA polymerase II in transcription. The CTD consists of multiple heptapeptide repeats; differences in the number of repeats between organisms and their degree of conservation have intrigued researchers for two decades. Here, we review the evolution of the CTD at the molecular level. Several primitive motifs have been integrated into compound heptads that can be readily amplified. The selection of phosphorylatable residues in the heptad repeat provided the opportunity for advanced gene regulation in eukaryotes. Current findings suggest that the CTD should be considered as a collection of continuous overlapping motifs as opposed to a specific functional unit defined by a heptad.**

## The evolution of RNA polymerases: new functions and a new domain?

The capacity of organisms to evolve is dependent on an ability to improve their functions and adapt to changing environments. Imperfections in the fidelity of DNA replication systems and environmental damage have altered the genome, with positive and detrimental results. Genes have been deleted, duplicated and inserted to yield genes with new functions. Higher eukaryotes can expand their repertoire of gene products further by alternative promoter usage and differential splicing to produce multiple transcripts from the same gene (for a review, see Ref. [1]). The evolution of RNA polymerases is no exception: prokaryotes produce one universal RNA polymerase, whereas three RNA polymerases (known as Pol I–III) perform distinct transcriptional functions in eukaryotes: Pol I transcribes rRNA; Pol II transcribes mRNA and Pol III transcribes tRNA. Despite differences in their coding sequence, all three polymerases contain the same structurally conserved active site [2]. Intriguingly, in contrast to other RNA polymerases, the largest subunit of RNA polymerase II (Rpb1) of higher organisms has developed a unique, repetitive structure at its C terminus. Biochemical and genetic studies have shown this domain to be essential for multiple steps in the regulation of gene expression, from the initiation of transcription on a chromatin template to the splicing and processing of the resulting RNA transcripts [3–6]. Prokaryotic RNA polymerase lacks an equivalent structure and cannot support RNA splicing. This Rpb1 C-terminal domain (CTD) is composed of heptad repeats, which would seem to have increased in number with organism complexity: mammalian Rpb1 has 52 heptad repeats, whereas *Plasmodium yoelii* possesses just 5 heptad repeats (Figures 1 and 2). This suggests that the repeat structure was (i) extended as a result of genetic instability and (ii) a critical development that has been under purifying selection, with the addition of more repeats increasing its functional efficiency. An examination of CTDs from different organisms not only reveals differences in length but also in heptad sequences [6]; nevertheless, a common structure of heptad repeats is retained across organisms (Figures 1 and 2). However, it was not clear how this sequence originated, and why heptad repeats? Was there a universal common ancestor? How can a common selection pressure have resulted in the large deviations in sequence and length observed between organisms?

In this review, we reconsider the origins and structure of the CTD in light of current molecular and genetic data. The obvious heptad-repeat structure, although pleasing to the eye, would appear to have distracted scientists from the real functional units of the CTD. Our examination suggests a new hypothesis for CTD evolution and allows us to explain both the high conservation and divergence within this enigmatic structure.

## Substructures within CTD heptads

In general, the CTD heptads consist of a common structure of tyrosine (or other hydrophobic amino acids) and two prolines, separated mostly by serines or other residues, producing the canonical heptad, tyrosine-serine-proline-threonine-serine-proline-serine (YSPTSPS). Heptad repeats, or CTD-like sequences, which comprise part of the heptad sequence, are also abundant in the C-terminal region downstream of domain H in Rpb1, commonly referred to as the Linker region (Figure 1a). These CTD-heptad submotifs generally take the form YSPx and SpxY, which, when overlapped and combined, produce a heptad with the register SPxYSPx (where x signifies any amino acid) (Figure 1b). For the purposes of this review, we used this register for the presentation of all CTD sequences. Using this form, the heptads, and submotifs thereof, are clearly identifiable in a variety of organisms (Figure 1c). Interestingly, the poorly described 'Linker' region is highly abundant in heptad submotifs. We have therefore redefined the CTD as all sequences C-terminal of Rpb1 domain H (Figure 1a).

On the basis of published CTD sequences [6,7], it is apparent that the CTD is divided into three distinct

*Corresponding authors:* Chapman, R.D. (drrobchapman@googlemail.com); Eick, D. (eick@helmholtz-muenchen.de).

0168-9525/$ – see front matter  doi:10.1016/j.tig.2008.03.010 Available online 9 May 2008

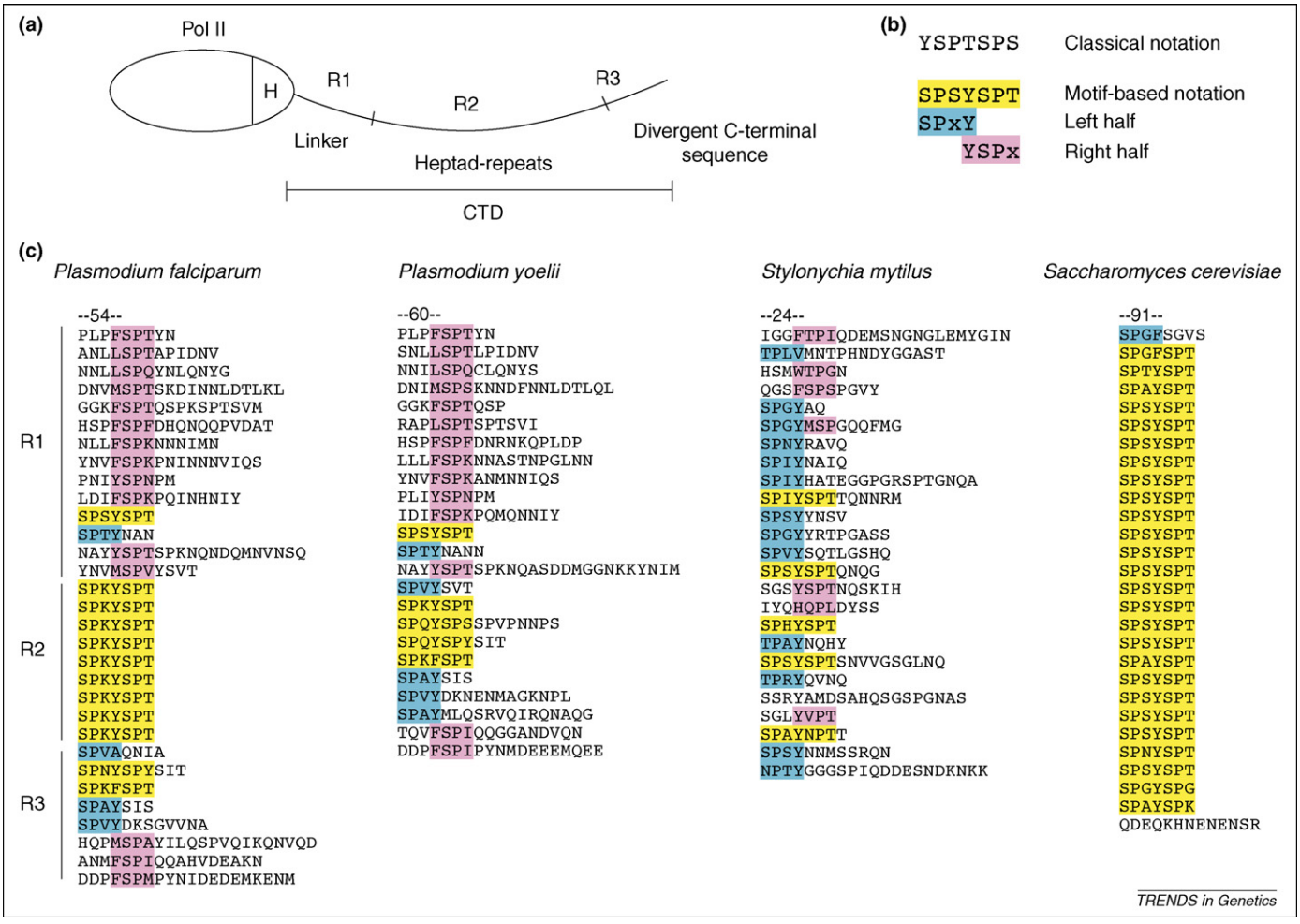


**Figure 1**. C-terminal domain (CTD) in simple organisms. **(a)** A simple representation of polymerase II (Pol II) with CTD organization. Domain H is highly conserved throughout organisms from bacteria to humans. The C-terminal region downstream of domain H often contains three distinct regions (R1–R3), as exemplified by *Plasmodium falciparum* [in **(c)**]. **(b)** Examination of the CTD in a variety of organisms revealed a mixture of sequences based around repeats of the classical, 'consensus' or 'canonical' heptad YSPTSPS. Historically, repeats of canonical and noncanonical heptads have been ordered into multiples of the YSPxSPx heptad. This has naturally led to the assumption that this heptad is the functional unit of the CTD. Recent data [7,8,50] suggested that this is not the case. We propose that the heptad structure resulted from the combination of submotifs YSPx and SPxY (where Y is frequently replaced by other hydrophobic amino acids) into a continuous overlapping sequence with shared residues. To visualize these motifs, we have therefore changed the heptad register of the sequences presented to SPxYSPx. **(c)** Differences in the regions C-terminal of domain H in *Plasmodium falciparum*, *Plasmodium yoelii*, *Saccharomyces cerevisiae* and *Stylonychia mytilus*. *S. cerevisiae* has a CTD with a large stretch of identical heptad repeats. It possesses few submotifs in its linker (not shown) in comparison to the other organisms shown. For most organisms, the regions R1 and R3 contain mainly single motifs on the right (pink) or left side (blue) of the central tyrosine/phenylalanine residue [see (b)]. The frequencies of these different motifs differ between organisms, with some being combined to form a compound heptad (yellow). *P. yoelii* and *P. falciparum* have a common origin; however, the latter contains an extended tract of eight SPKYSPT heptads. The number of amino acids between the H-domain and the start of sequences is shown above. GenBank accessions used: *P. yoelii*, XM_726075; *P. falciparum*, Z98551; *S. mytilus*, AF315823; *S. cerevisiae*, Z74188.

regions: Linker and heptad-related sequences (R1); a heptad-repeat region (R2) and divergent C-terminal sequences (R3) (Figure 1a). The size and occurrence of each of these regions differs between organisms (Figure 1c). A few heptads, and an abundance of submotifs, can be identified in organisms not previously thought to have developed a significant CTD: *Stylonychia mytilus*, a ciliate, exhibits more of the subconsensus motifs SPxY than *Plasmodium falciparum* or *P. yoelii*, which have more YSPx motifs. When combined, these motifs form the compound heptad motif SPxYSPx, or alternatively the octad YSPxSPxY, implying that the functional unit of CTD motifs might cross into neighbouring repeats. Indeed, this would explain the observation that yeast, which has a CTD with many repeats of an identical heptad (Figure 1c), can survive with a CTD composed of di-heptads separated by spacers (a few alanine residues) but not spaced mono-heptads [8]. A recent study by the same group has independently refined the essential sequence elements required for CTD function in yeast to the same YSPxSPxY octad; however, this is only in combination with a proximal SPxSPx or a distal SP [9]. These data therefore further support our hypothesis that CTD function arose from the fusion of submotifs into a continuous overlapping structure. The heptad-repeat region (R2) of many CTDs is further divided into separate regions that contain identical heptads and heterogeneous heptads, as exemplified by the proximal and distal regions of human CTD (Figure 2). Finally, the divergent sequences at the carboxy terminus (R3) have little conservation between organisms. In mammals, part of this sequence is phosphorylated by CKII [10], binds to the Abl1/2 tyrosine kinases [11,12] and is required to prevent CTD degradation [10,13]. It does not seem to have a specific role in RNA processing as previously suggested [14] because it can be replaced by random sequences without any affect on function [13]. Early

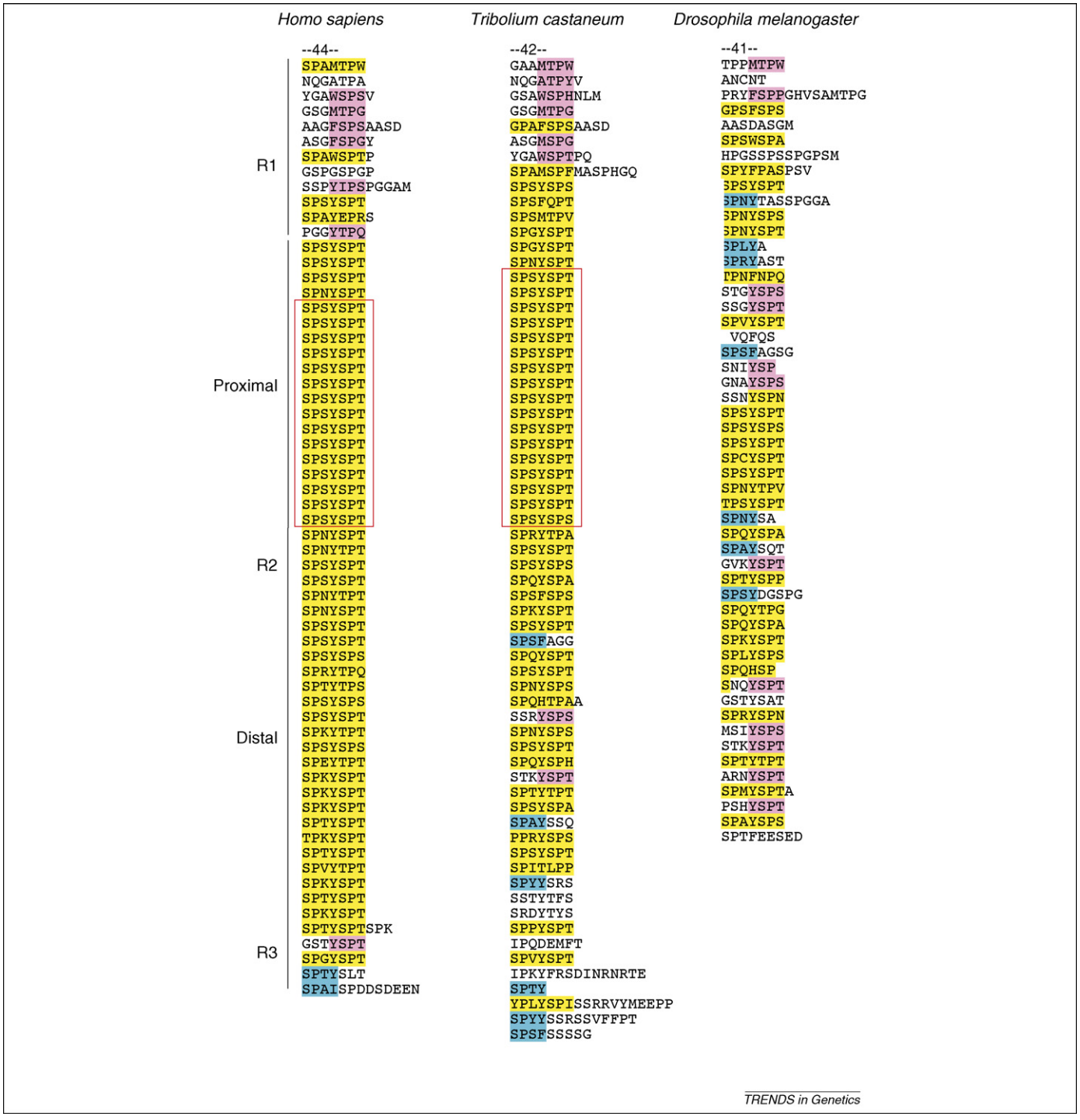


**Figure 2**. Comparison of organisms with extended heptad C-terminal domains (CTDs). The Rpb1 CTDs of *Homo sapiens* (human) and *Tribolium castaneum* (red flour beetle) have large tracts of the canonical heptad SPSYSPT (red boxes). By contrast, the *Drosophila melanogaster* (fruit fly) CTD contains fewer heptads and exhibits little homology between those it has. It is devoid of tracts of identical heptads. The differences in these sequences could be explained by independent expansion of CTD heptads or as the result of degeneration from a common ancestor CTD precursor. GenBank accessions used: *H. sapiens,* NM_000937; *T. castaneum,* XM_968377; *D. melanogaster,* NM_078569.

experiments failed to account for CTD degradation in transcription experiments. These differences in CTD sequence suggest that changing the heptad sequence composition might affect CTD function.

## The composition of heptad repeats

Given the high conservation of the CTD heptad-repeat structure, the differences in the heptad sequences observed is perhaps surprising: although the tyrosine and prolines maintain the repetitive structure, the flanking residues in between vary. The relevance of these deviations is unclear. The CTD of *Drosophila melanogaster* Rpb1 (Figure 2) contains few heptads with the same sequence, whereas the heptads of mammalian CTDs are mostly homologous, except for the deviations at position 7. *Drosophila's* CTD seems to be an anomaly, because large tracts of identical heptads can be seen in other organisms including the distantly related *Tribolium castaneum* (the red flour

beetle; Figure 2); the CTDs of the yeast *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* are almost identical throughout (SPSYSPT) [6,15,16], as are the 25 repeats of the protist *Mastigoamoeba invertens* (SPAYSPA) [17] and the 9 repeats of *P. falciparum* (SPKYSPT) [18] (Figure 1c). What sets these organisms apart, however, is their choice of heptad. One group of organisms, known as the CTD clade [19], is defined by its high conservation of the canonical heptad SPSYSPT. This particular heptad is rich in residues that can be phosphorylated, thereby enabling the possibility of regulation. An expansion involving identical heptads might indicate a recent amplification event [20], but their high degree of conservation also suggests that they were positively selected. Could the similarity between CTDs in such organisms be explained through the common selection of an ancestral heptad? [21,22].

## The DNA sequence reveals the history of heptad amplifications

An examination of the DNA sequences encoding the CTD region can reveal the nature of the heptad expansion in different species [20]. Because of the degenerate nature of the genetic code, most amino acids can be specified by several codons. If we assume that the CTD developed through amplification, one would expect codon usage to be conserved between repeats. Heterogeneous codon usage would therefore suggest an independent origin of the CTD or indicate older duplications that have had time to experience mutation-causing damage. Serine is present in three positions in the canonical CTD clade heptad ($S_5PS_7YS_2PT$) and can be encoded by six codons (i.e. 4× TCx; 2× AGc/u). Interestingly, large stretches of sequence or entire CTDs are composed of repeats using specific codon constellations (i.e. specific patterns of codon usage) across the three serine positions ($S_5$, $S_7$ and $S_2$) (Figure 3a) [17]. Multiples of heptads with the same codon constellation imply expansion from a common heptad [17,20]; however, these constellations used differ greatly between CTD clade organisms, as exemplified by the differences between *S. cerevisiae*, *S. pombe* and *Variamorpha necatrix* (Figure 3a). This suggests that identical tracts of heptad-repeats have different evolutionary origins. It is highly unlikely that these defined patterns would result from degeneration. One such tract in the proximal region of mammalian CTD (Figure 3b, red box) exclusively uses the codons TCx (ser5), AGc/u (ser7) and TCx (ser2) but exhibits frequent differences at the third codon, or 'wobble' position. Although there is a significant number of mutations at this position, the lack of changeover between the serine AGc/u and TCx codons suggests that CTD degeneration is not the major source of the differences observed in CTD coding

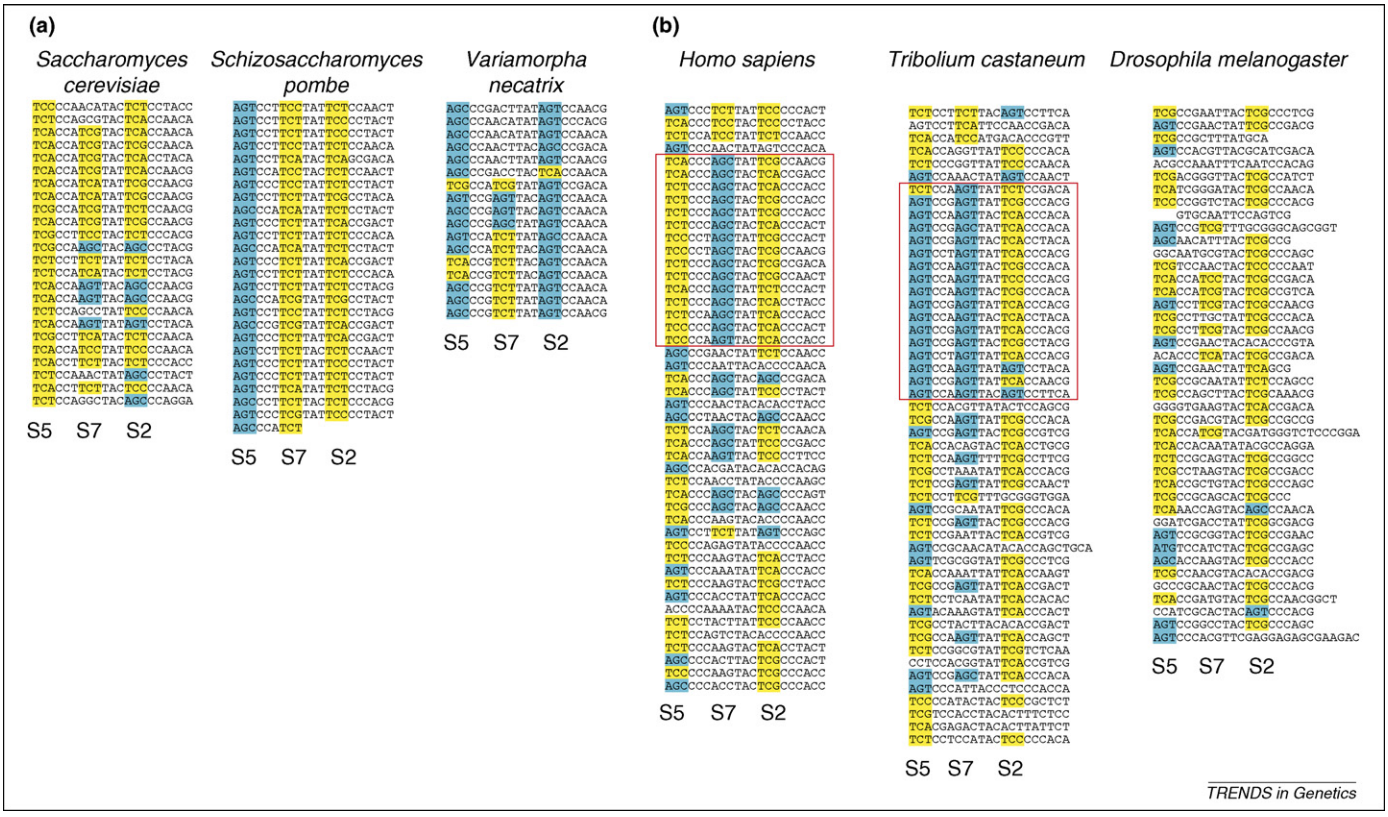


**Figure 3**. Analysis of codon usage in the C-terminal domain (CTD) region. **(a)** The amino acid serine has six possible codons (TCx and AGc/t). Examination of their distribution within the CTD of three different organisms reveals a remarkable conservation at positions within the heptad (S2, S5, S7). Importantly, each organism shows a different constellation of codons across these positions. This implies that the CTD has evolved independently in most organisms, through amplification of heptads and degeneration of amplified regions. The lack of degeneration in *Saccharomyces pombe* suggests it was recently amplified from just one precursor heptad or has lost its degenerate heptads. In this respect, it is unclear whether the CTDs observed here have replaced an ancient predecessor. GeneBank accessions used: *Saccharomyces cerevisiae*, Z74188; *S. pombe*, NM_001021568; *Variamorpha necatrix*, AF060234. **(b)** Analysis of codon usage in the extended CTDs from more complex organisms reveals different degrees of conservation between regions. This is highest in the sequences coding for tracts of identical heptads (red box), indicating that they result from expansion of the same sequence. The pattern is more random in the sequences outside this tract. This suggests that such CTDs evolved in a stepwise manner before a sequence of SPSYSPT heptads emerged that could be readily amplified. The sequence of *Drosophila melanogaster* does not possess a tract of identical heptads.

sequences among organisms [23] (Figure 3b); more likely, it is the result of differences in the genetic shuffling of ancestral sequences over time.

The homogeneity of codon-constellations in individual species and the heterogeneity seen between them suggest that our common CTD ancestor might not have possessed a canonical repeat but instead had a serine/proline-rich precursor sequence, similar to that in the Linker region (R1). Multiples of near-canonical heptads are found in several organisms thought to be unrelated to organisms of the CTD clade [17,20]. In addition, several proteins, for example, the zinc-finger protein 768 (ZF768), WW domain-binding protein 2 (WBP2-like) and post-acrosomal sheath WW domain-binding protein (PAWP), have a similar heptad repeat structure (Figure I of Box 1). The expansion of such sequences suggests that they have some important selective advantage.

The DNA sequence of mammalian CTD could imply an original selection for the imperfect repeats that are contained in its distal region, where serine7 is replaced by lysine. There is little conformity in the codon-usage of these heptads, suggesting that either they have degenerated or have independent origins; however, their conservation among mammals suggests they are functional. It is possible that such heptad repeats were a predecessor to mammalian CTD and appeared before the expansion of the canonical heptads. Although little is known of their function, these noncanonical heptads are the preferred substrates of the kinases CDK1/cdc2 *in vitro* [24] and do not appear equivalent to their canonical counterparts *in vivo* [13].

## The importance of heptad sequence and CTD length

The presence of CTD-heptads in such diverse organisms as amoeba [17], plasmodium and mammals [6] could indicate that they facilitate an interaction with an evolutionarily old and highly conserved structure. CTD has previously been shown to bind DNA [25], possibly with the purpose of displacing negative regulators, thereby facilitating interactions between activators and the transcription factor at promoters [6], but whether this is an function of CTD *in vivo* remains to be seen. CTDs repetitive nature suggests that it functions by increasing the number of available interactions for CTD-binding proteins; however, investigations into the importance of CTD length, in yeast and mammals, have not yet explained why its length is so highly conserved in a given species, because viable yeast and mice can be produced with truncated forms [26,27].

Initial studies into CTD function were performed in mammalian cell lines but proved limited in their usefulness, given the lack of analytical tools available at that time. In the 1990s, research moved to yeast, given its

**Box 1. Proteins with C-terminal domain–related structures**

Heptad-repeat structures can also be seen in several other proteins, and like the polymerase II C-terminal domain (CTD), their functions are not clear (Figure I). One such protein, a zinc-finger protein (ZF768), possesses heptads with a CTD-like configuration [55] (SPx-F/Y-xPx), whereas a protein similar to WW domain binding protein 2 (WBP2-like) and protein acrosomal sheath WW domain-binding protein (PAWP) display a different heptad configuration (PxxYxxP). The latter have been reported to bind WW domains [56], which are also present in some CTD-binding proteins [4]. It is possible that such repetitive structures have been selected for their abilities to bind several WW domains at once. GenBank accession used: ZF768, AAH13760; WBP2-like, XP_001077583; PAWP, AK129656.

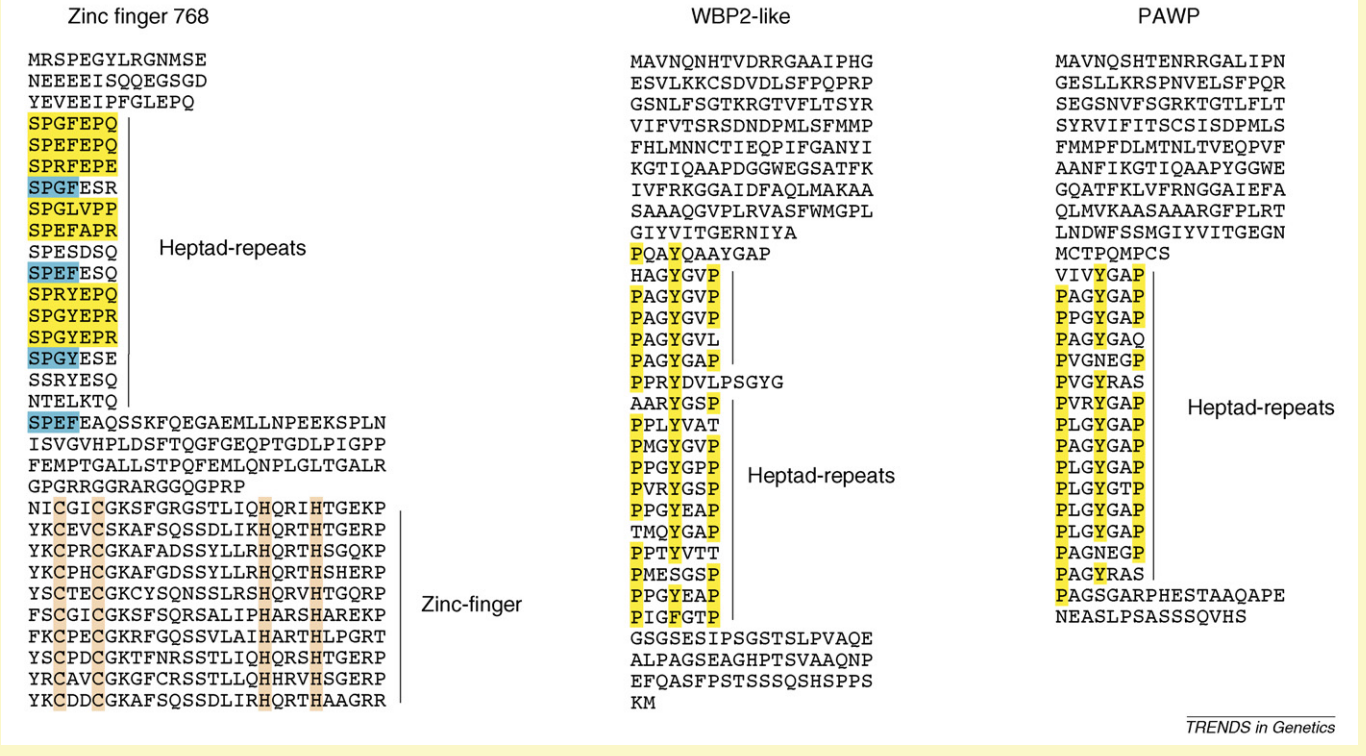


**Figure I**. The C-terminal domain of three proteins: ZF768, WBP2-like, and PAWP.

relative ease for use in genetic complementation studies. One such systematic investigation of yeast CTD length and composition by the Corden laboratory [27] produced viable strains with just eight SPSYSPT heptads, although such mutants exhibited weaker growth and sensitivity to environmental stress. This study also assessed the effect of noncanonical heptads and their positioning within the CTD. Surprisingly, there was a difference in the heptads that could be tolerated proximal or distal to the Linker region (R1): glutamic acid proved lethal in place of serine2 when proximal but not distal, whereas the opposite is true for similar replacement of serine5. This suggests subtle differences in function of regions of an otherwise homogenous sequence and has long been speculated for the mammalian CTD: most proximal repeats are canonical compared with just a few in the distal region. Research into CTD function in yeast is, however, complicated by the fact that some laboratory strains of *S. cerevisiae* have different lengths of CTD [16] and have the ability to overcome CTD truncation by amplifying heptads to create a new CTD.

The past decade has seen a return to mammalian cell systems: Fong and Bentley [28] demonstrated differences in the binding of the 3′-RNA processing enzyme, CstF, and capping enzyme to different CTD segments *in vitro* independent of CTD length; furthermore, replacement of certain noncanonical heptads with canonical heptads in mammalian CTD induces Rpb1 degradation [10,13,29]. Nevertheless, as the following studies suggest, CTD function seems to largely be dependent on its length, indiscriminate of sequence. In-depth investigation into the role of CTD in pre-mRNA 3′ cleavage indicated that, although wild-type mammalian CTD was a little more effective than an all-canonical CTD, activity was mostly dependent on CTD length [30]. Recently, it was shown that a mammalian Rpb1 with a CTD truncated to 31 repeats retains mRNA at the site of transcription, independent of splicing and 3′ processing [31]. A certain CTD length might be a requirement for the binding of RNA processing factors, as shown for PSF and $p54^{nrb}$/NonO [32]. Results from a two-hybrid assay suggested that a minimum length of 12 and 14 repeats was required for proper interaction with the yeast capping enzymes Pct1 and Pce1, respectively [33]; thus, the CTD length requirement is dependent on the protein with which the CTD interacts. Current data suggest that as few as 19–22 repeats are required for splicing and 3′ end cleavage functions in mammalian cell systems [29,34], which is also in line with our own findings that at least 16 canonical repeats are required for Rpb1 to support its own expression [35].

Complementation experiments in yeast revealed that certain genes, when deleted, could compensate for the lethal effects of truncating the CTD [36]. These genes, termed SRBs (suppressor of Rpb1 mutation), encode proteins that form part of complexes (e.g. the mediator) that are required for the initiation of transcription, but also exert a negative influence on transcription elongation [37].

The CTD is modified during the transcription cycle: Pol II with a nonphosphorylated CTD is recruited to genes [38]; its transition into the transcription elongation phase is regulated through phosphorylation of serines in the CTD, which release it from complexes such as Mediator [39], negative elongation factor (NELF) and DRB sensitivity–inducing factor (DSIF) [40]—factors that have a negative influence on transcription elongation. It is possible that a certain length of CTD is required to be sufficiently phosphorylated to protect it from the negative effects of such complexes. This idea is supported by complementation studies showing that deletion of Med13(Srb9), a component of the Mediator complex, rescues a lethal mutant of the CTD phospho-acceptor serine2 (ser2ala) [41].

## Canonical vs noncanonical repeats

In mammalian cell lines, Rpb1 mutants that lack CTD or that are composed only of certain noncanonical repeats, irrelevant of heptad repeat length, cannot support cell viability [13,35,42]. Such mutants can bind to promoters but seem to experience problems at subsequent stages of elongation [35,43,44]. How can this be explained? Certain nonconsensus sequences might impede binding of initiating complexes (e.g. Mediator), thereby reducing the chance of successful initiation. Noncanonical heptads often lack certain modifiable residues and are poorer substrates for CTD kinases [9,35,44].

## Differential requirement for canonical repeats between yeast and human

Despite the high conservation of the canonical SPSYSPT heptad across the CTD clade, there are different functional requirements for individual residues. CTDs composed of certain nonconsensus repeats can be tolerated in yeast but are lethal in mammals: Stiller *et al.* [17] could functionally replace the canonical CTD of *S. cerevisiae* with the 25-repeat ($SPA_7YSPA_4$) CTD of *M. invertans*, where threonine4 and serine7 are replaced by alanine throughout. Although it is possible to replace yeast CTD with its larger, mammalian counterpart and maintain function [15], replacement of mammalian CTD with a pure-canonical CTD of the same length, while tolerated, reduces cell viability [13]. The differential requirements for such residues might reflect a further development in control, because nonphosphorylatable residues might serve to limit phosphorylation [35] or provide another form of signalling, for example, through the acetylation or methylation of lysine.

## Signalling to CTD

The evolution of a canonical CTD clade of organisms has been accompanied by the appearance of the CTD kinases [22] and CTD-associated proteins [21]. The abundance of phosphorylatable residues in the canonical heptad endows it with great potential as an array for signal transduction. However, the purpose of this phosphorylation remained enigmatic, and it was only with the development of new techniques that the first evidence of CTD regulation through phosphorylation could be shown. However, this was confused by the finding that evolutionary-related kinases are not equivalent in regulation of CTD function in all organisms. This might therefore explain differences in CTD heptad sequence, because the CTD evolved alongside the substrate preference of CTD kinases [9,22]. In 2000, results from Buratowski's laboratory showed that the differential phosphorylation of serines 2 and 5 during

different phases of transcription coincided with the presence of different mRNA processing factors [45], leading to the idea of a 'CTD code' that regulated Pol II transcription and mRNA processing [46,47] (For a recent update on the CTD code, see the accompanying article in this issue by Egloff *et al.* [48]). A variety of kinases have now been shown to phosphorylate CTD, in addition to a variety of proteins that bind phospho-CTD [4,5,48]. Recently, it has been shown that phosphorylation of serine7 in the canonical heptad [35] regulates snRNA processing in mammals [35,49]. This poses interesting questions as to the conservation of CTD functions, because this position exhibits the most deviation between organisms of the CTD clade. It is highly likely that, in addition to core functions, individual organisms have evolved further unique CTD functions. This might be reflected in the heptad motifs recognized by core factors and those recognized by species-specific factors.

## CTD-binding factors

Current structural data suggest different CTD-binding proteins have specific requirements for interacting with CTD, including its phosphorylation status. The CTD-interacting domains (CIDs) of the 3′-RNA processing factors Pcf11 and Nrd1 require phosphorylation of the second serine within the sequence PSYSPTSP for binding [50]. The peptidyl-prolyl isomerase Pin1 interacts with the sequence SPTSPS within one repeat [51], whereas two repeats are required for interaction with the yeast capping enzyme Cgt1 [52] (requiring YSPTS in each, where the fifth residue, serine, is phosphorylated), and the histone H3 methylase Set2 requires tyrosines from two adjacent repeats [53,54] (for a detailed overview, see Egloff *et al.* [48] and [9]). It is notable that, with the exception of CTD-modifying enzymes, all the CIDs identified thus far have a requirement for residues across more than one canonical heptad. These findings again imply that the heptad structure should be considered as a set of overlapping motifs.

## Concluding remarks

From our assessment of the available data, it would seem that heptad-rich C-terminal domain (CTD)-like structures have arisen separately in different organisms several times during the course of evolution. The argument for convergence is supported both by the variety of heptad sequences observed and their underlying DNA coding sequence. The universal selection of heptads, as opposed to other sequences, suggests an initial common purifying selection pressure. However, the type of heptad originally formed might have defined CTD function for an organism's future evolution. For example, *Drosophila melanogaster's* varied CTD suggests it evolved within parameters that were different to those of the organisms within the conserved canonical CTD clade, because *Drosophila* does not seem to have developed a dependence on tracts of canonical repeats. It is therefore unlikely that all CTDs have exactly the same functional characteristics or that there is a universal CTD code. It is plausible that, although certain basic signals are conserved, some heptads possess advanced functions that are unique to a given species. The recent revelation that phosphorylation of a specific residue of the mammalian CTD canonical heptad regulates snRNA processing lends support to this theory, because this position is often substituted for a non–phospho-acceptor in many other CTD clade organisms. The development of analytical techniques, such as DNA-microarray analysis of chromatin immunoprecipitates (ChIP on chip), and antibodies to examine CTD phosphorylation status should now enable scientists to determine whether there is indeed a CTD code across genomes. Box 2 lists some areas for future research. In light of structural data, it is now important to reconsider the role of individual submotifs and their functions within both the Linker region and in organisms previously considered to have no CTD-like structure.

**Box 2. Questions for future research**

1) What has been the driving force for the development of the overlapping heptad-repeat structure?
2) What groups of factors can interact with the C-terminal domain (CTD)? In addition to proteins, DNA has also been shown to bind the CTD through regularly spaced intercalation of aromatic groups from tyrosine within DNA base pairs. Could DNA or RNA have been the purifying selection pressure for CTD's regular structure?
3) CTD is essential for transcription on chromatin templates and has been shown to bind the histone-modifying enzymes Set1 and Set2 (for a review, see Ref. [4]). Furthermore, CTD is only present in organisms that use chromatin to pack their DNA. Have CTD and chromatin co-evolved? Can the CTD interact directly with nucleosomes? Can epigenetic markers in chromatin affect CTD modification?
4) CTD is not required for the enzymatic process of RNA transcription but is essential for its regulation and post-transcriptional processes such as the splicing and 3′-processing of RNA. Has the repetitive CTD structure been selected for simultaneous binding of multiple intron–exon junctions?
5) What are the differences between the different CTD submotifs? Do certain heptads (i.e. noncanonical repeats) have specific functions?
6) The recent identification of a new signalling site in mammalian CTD invites speculation as to whether threonine and tyrosine phosphorylation [57] in CTD motifs regulate polymerase II function.

**Acknowledgements**

The authors thank Shona Murphy, Patrick Cramer, Stefan Boeing, Thomas Albert and Marcus Conrad for useful discussions and careful reading of the manuscript. Work in our laboratory is supported by grants from the Deutsche Forschungsgemeinschaft (SFB/TR5), Boehringer Ingelheim Fonds and Fonds der Chemischen Industrie.

## References

1 Blencowe, B.J. (2006) Alternative splicing: new insights from global analyses. *Cell* 126, 37–47
2 Cramer, P. *et al.* (2001) Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* 292, 1863–1876
3 Sims, R.J., 3rd *et al.* (2004) Elongation by RNA polymerase II: the short and long of it. *Genes Dev.* 18, 2437–2468
4 Phatnani, H.P. and Greenleaf, A.L. (2006) Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev.* 20, 2922–2936
5 Palancade, B. and Bensaude, O. (2003) Investigating RNA polymerase II carboxyl-terminal domain (CTD) phosphorylation. *Eur. J. Biochem.* 270, 3859–3870
6 Corden, J.L. and Ingles, C.J. (1992) Carboxy-terminal domain of the largest subunit of eukaryotic RNA Polymerase II. In *Transcriptional Regulation* (McKnight, S.L. and Yamamoto, K.R., eds), pp. 81–108, Cold Spring Harbor Laboratory Press

7 Chapman, A.B. and Agabian, N. (1994) Trypanosoma brucei RNA polymerase II is phosphorylated in the absence of carboxyl-terminal domain heptapeptide repeats. *J. Biol. Chem.* 269, 4754–4760
8 Stiller, J.W. and Cook, M.S. (2004) Functional unit of the RNA polymerase II C-terminal domain lies within heptapeptide pairs. *Eukaryot. Cell* 3, 735–740
9 Liu, P. *et al.* (2008) The essential sequence elements required for RNAP II carboxyl-terminal domain in yeast and their evolutionary conservation. *Mol Biol. Evol.* 25, 719–727
10 Chapman, R.D. *et al.* (2004) The last CTD repeat of the mammalian RNA polymerase II large subunit is important for its stability. *Nucleic Acids Res.* 32, 35–44
11 Baskaran, R. *et al.* (1997) Tyrosine phosphorylation of RNA polymerase II carboxyl-terminal domain by the Abl-related gene product. *J. Biol. Chem.* 272, 18905–18909
12 Baskaran, R. *et al.* (1999) Nuclear c-Abl is a COOH-terminal repeated domain (CTD)-tyrosine (CTD)- tyrosine kinase-specific for the mammalian RNA polymerase II: possible role in transcription elongation. *Cell Growth Differ.* 10, 387–396
13 Chapman, R.D. *et al.* (2005) Role of the mammalian RNA polymerase II C-terminal domain (CTD) nonconsensus repeats in CTD stability and cell proliferation. *Mol. Cell. Biol.* 25, 7665–7674
14 Fong, N. *et al.* (2003) A 10 residue motif at the C-terminus of the RNA pol II CTD is required for transcription, splicing and 3′ end processing. *EMBO J.* 22, 4274–4282
15 Allison, L.A. *et al.* (1988) The C-terminal domain of the largest subunit of RNA polymerase II of Saccharomyces cerevisiae, *Drosophila melanogaster*, and mammals: a conserved structure with an essential function. *Mol. Cell. Biol.* 8, 321–329
16 Nonet, M. *et al.* (1987) Functional redundancy and structural polymorphism in the large subunit of RNA polymerase II. *Cell* 50, 909–915
17 Stiller, J.W. *et al.* (1998) Amitochondriate amoebae and the evolution of DNA-dependent RNA polymerase II. *Proc. Natl. Acad. Sci. U. S. A.* 95, 11769–11774
18 Li, W.B. *et al.* (1989) An enlarged largest subunit of Plasmodium falciparum RNA polymerase II defines conserved and variable RNA polymerase domains. *Nucleic Acids Res.* 17, 9621–9636
19 Stiller, J.W. and Hall, B.D. (2002) Evolution of the RNA polymerase II C-terminal domain. *Proc. Natl. Acad. Sci. U. S. A.* 99, 6091–6096
20 Giesecke, H. *et al.* (1991) The C-terminal domain of RNA polymerase II of the malaria parasite Plasmodium berghei. *Biochem. Biophys. Res. Commun.* 180, 1350–1355
21 Guo, Z. and Stiller, J.W. (2005) Comparative genomics and evolution of proteins associated with RNA polymerase II C-terminal domain. *Mol. Biol. Evol.* 22, 2166–2178
22 Guo, Z. and Stiller, J.W. (2004) Comparative genomics of cyclin-dependent kinases suggest co-evolution of the RNAP II C-terminal domain and CTD-directed CDKs. *BMC Genomics* 5, 69
23 Stiller, J.W. *et al.* (2000) Evolutionary complementation for polymerase II CTD function. *Yeast* 16, 57–64
24 Rickert, P. *et al.* (1999) Cyclin C/CDK8 and cyclin H/CDK7/p36 are biochemically distinct CTD kinases. *Oncogene* 18, 1093–1102
25 Suzuki, M. (1990) The heptad repeat in the largest subunit of RNA polymerase II binds by intercalating into DNA. *Nature* 344, 562–565
26 Litingtung, Y. *et al.* (1999) Growth retardation and neonatal lethality in mice with a homozygous deletion in the C-terminal domain of RNA polymerase II. *Mol. Gen. Genet.* 261, 100–105
27 West, M.L. and Corden, J.L. (1995) Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations. *Genetics* 140, 1223–1233
28 Fong, N. and Bentley, D.L. (2001) Capping, splicing, and 3′ processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD. *Genes Dev.* 15, 1783–1795
29 de la Mata, M. and Kornblihtt, A.R. (2006) RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. *Nat. Struct. Mol. Biol.* 13, 973–980
30 Ryan, K. *et al.* (2002) Requirements of the RNA polymerase II C-terminal domain for reconstituting pre-mRNA 3′ cleavage. *Mol. Cell. Biol.* 22, 1684–1692
31 Custodio, N. *et al.* (2007) Splicing- and cleavage-independent requirement of RNA polymerase II CTD for mRNA release from the transcription site. *J. Cell Biol.* 179, 199–207
32 Rosonina, E. *et al.* (2005) Role for PSF in mediating transcriptional activator-dependent stimulation of pre-mRNA processing *in vivo*. *Mol. Cell. Biol.* 25, 6734–6746
33 Pei, Y. *et al.* (2001) The length, phosphorylation state, and primary structure of the RNA polymerase II carboxyl-terminal domain dictate interactions with mRNA capping enzymes. *J. Biol. Chem.* 276, 28075–28082
34 Rosonina, E. and Blencowe, B.J. (2004) Analysis of the requirement for RNA polymerase II CTD heptapeptide repeats in pre-mRNA splicing and 3′-end cleavage. *RNA* 10, 581–589
35 Chapman, R.D. *et al.* (2007) Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. *Science* 318, 1780–1782
36 Thompson, C.M. *et al.* (1993) A multisubunit complex associated with the RNA polymerase II CTD and TATA-binding protein in yeast. *Cell* 73, 1361–1375
37 Myers, L.C. and Kornberg, R.D. (2000) Mediator of transcriptional regulation. *Annu. Rev. Biochem.* 69, 729–749
38 Lu, H. *et al.* (1991) The nonphosphorylated form of RNA polymerase II preferentially associates with the preinitiation complex. *Proc. Natl. Acad. Sci. U. S. A.* 88, 10004–10008
39 Max, T. *et al.* (2007) Hyperphosphorylation of the C-terminal repeat domain of RNA polymerase II facilitates dissociation of its complex with mediator. *J. Biol. Chem.* 282, 14113–14120
40 Peterlin, B.M. and Price, D.H. (2006) Controlling the elongation phase of transcription with P-TEFb. *Mol. Cell* 23, 297–305
41 Yuryev, A. and Corden, J.L. (1996) Suppression analysis reveals a functional difference between the serines in positions two and five in the consensus sequence of the C-terminal domain of yeast RNA polymerase II. *Genetics* 143, 661–671
42 Bartolomei, M.S. *et al.* (1988) Genetic analysis of the repetitive carboxyl-terminal domain of the largest subunit of mouse RNA polymerase II. *Mol. Cell. Biol.* 8, 330–339
43 Meininghaus, M. *et al.* (2000) Conditional expression of RNA polymerase II in mammalian cells. Deletion of the carboxyl-terminal domain of the large subunit affects early steps in transcription. *J. Biol. Chem.* 275, 24375–24382
44 Lux, C. *et al.* (2005) Transition from initiation to promoter proximal pausing requires the CTD of RNA polymerase II. *Nucleic Acids Res.* 33, 5139–5144
45 Komarnitsky, P. *et al.* (2000) Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev.* 14, 2452–2460
46 Buratowski, S. (2003) The CTD code. *Nat. Struct. Biol.* 10, 679–680
47 Corden, J.L. (2007) Transcription. Seven ups the code. *Science* 318, 1735–1736
48 Egloff, S. and Murphy, S. Deciphering the expanding RNA polymerase II CTD code. *Trends Genet*. (in press)
49 Egloff, S. *et al.* (2007) Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. *Science* 318, 1777–1779
50 Meinhart, A. and Cramer, P. (2004) Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA-processing factors. *Nature* 430, 223–226
51 Verdecia, M.A. *et al.* (2000) Structural basis for phosphoserine-proline recognition by group IV WW domains. *Nat. Struct. Biol.* 7, 639–643
52 Fabrega, C. *et al.* (2003) Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Mol. Cell* 11, 1549–1561
53 Li, M. *et al.* (2005) Solution structure of the Set2-Rpb1 interacting domain of human Set2 and its interaction with the hyperphosphorylated C-terminal domain of Rpb1. *Proc. Natl. Acad. Sci. U. S. A.* 102, 17636–17641
54 Vojnic, E. *et al.* (2005) Structure and CTD binding of the Set2 SRI domain that couples histone H3 K36 methylation to transcription. *J. Biol. Chem.*, DOI: 10.1074/jbc.C500423200
55 Strausberg, R.L. *et al.* (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci. U. S. A.* 99, 16899–16903
56 Wu, A.T. *et al.* (2007) PAWP, a sperm-specific WW domain-binding protein, promotes meiotic resumption and pronuclear development during fertilization. *J. Biol. Chem.* 282, 12164–12175
57 Baskaran, R. *et al.* (1993) Tyrosine phosphorylation of mammalian RNA polymerase II carboxyl- terminal domain. *Proc. Natl. Acad. Sci. U. S. A.* 90, 11167–11171